CDF/DOC/CDF/PUBLIC/10803
Version 1.1
March 12, 2012

Introduction to HOBIT, a b Jet Identification Tagger at the CDF Experiment Optimized for Light Higgs Searches

Thomas Phillips[1]

Duke University

John Freeman[2] Thomas Junk[3] Mike Kirby[4] Rick Snider[5]

Fermilab

Weiming Yao[6]

LBNL

Jesus Vizan[7]

Universidad de Oviedo

Marco Trovato[8]

Piza

Yuri Oksuzian[9]

University of Virginia

Abstract

We present the development and validation of the Higgs-Optimized b-Identification Tagger (HOBIT), a multivariate b jet identification algorithm optimized for Higgs boson searches at the Fermilab Tevatron's CDF experiment. At collider experiments, b taggers allow one to distinguish particle jets containing B hadrons from other jets; these algorithms have been used for many years with great success at CDF. HOBIT is unique among CDF b taggers both in the extent to which it synthesizes and extends the best ideas of previous taggers, as well as the fact that it has been specially designed to work best in searches for $H \to b\bar{b}$ decay, as opposed to being an all-purpose tagger. Employing feed-forward neural network architectures, HOBIT provides an output value ranging from approximately -1 ("light jet like") to 1 ("b jet like"); this continuous output value has been tuned on by light Higgs search analyses so as to provide maximum sensitivity. Along with the features of the tagger, its characterization in the form of b jet finding efficiencies and false (light jet) tag rates is presented.

[1]thomas.phillips@duke.edu
[2]jcfree@fnal.gov
[3]trj@fnal.gov
[4]kirby@fnal.gov
[5]rs@fnal.gov
[6]weiming@fnal.gov
[7]vizan@fnal.gov
[8]mtrovato@fnal.gov
[9]oksuzian@fnal.gov

Contents

1 Introduction

1.1 Context

The identification of jets originating from b quarks is an important part of many analyses at high-energy physics colliders, including the study of top physics, searches for beyond-the-standard-model phenomena, and searches for a light Higgs boson undergoing a $H \to b\bar{b}$ decay. At CDF, the search for a light Higgs boson has been a subject of increasing interest and focus in recent years. While there have been numerous successful b taggers over the years, most have essentially been intended to serve as "general-purpose" taggers, that is, their efficiencies and purities, as well as the types of b jets/non-b jets they are designed to accept/reject, have been chosen without a specific analysis in mind. However, aspects of a given analysis, such as the optimal signal-to-background ratio, or the relative proportion of non-b jets originating from gluons, can influence whether a tagger can be considered optimal for the context in question. Due to the relatively low cross section of Higgs production at Tevatron energies, traditional taggers have tended toward a higher purity and lower efficiency than would be ideal for Higgs searches. While this problem can be circumvented somewhat by the logical OR-ing of different taggers' acceptances, a more elegant and flexible solution can be

found in the tunability inherent in the continous output of a neural network output such as HOBIT provides.

1.2 Physics of b Decay

The salient features of a jet of particles resulting from a b quark decay are due to the production of a B hadron whose relatively long lifetime allows it to travel a macroscopic distance before decaying. Here, "macroscopic" is meant to describe a distance on the order of a millimeter, given that, ignoring relativistic effects, the mean decay length of a B^0 ($B^{\pm}$, Λ_b) hadron is 460 μm (501 μm, 367 μm); these distances are increased by the time dilation of the hadron's decay brought on by its highly relativistic velocity. These displacements between the location of the $p\bar{p}$ collision (the primary vertex) and the B hadron decay (the secondary vertex) can be resolvable by the CDF tracking system, in particular its silicon detector. Almost all information as to whether or not a given jet originates from a B hadron decay is carried in the tracks its charged particles leave in the detector. Specifically, it is possible to identify the delayed decay of a B hadron through the displacement of the tracks of individual charged particles from the hadron decay with respect to the primary vertex and also through the combining of tracks in the form of a fitted secondary decay vertex.

Other features also distinguish the b jet. Due to the large mass of the b quark, the decay products of B hadrons will form a larger invariant mass than those of hadrons not containing b quarks. Furthermore, the large relativistic boost typical of a B hadron will result in decay products which tend to be more energetic and collimated within a jet cone than other particles. Finally, particle multiplicities tend to be different for jets containing B hadron decays compared to other jets; in particular, muons and electrons appear in approximately 20% of jets containing a B hadron, typically either directly via semileptonic decay of the B or indirectly through the semileptonic decay of a D or Λ_c resulting from a B decay.

2 b Tagging Algorithms

As a tremendous amount of effort has gone into the construction of b taggers at CDF and other experiments [1, 2, 3] , it made little sense to disregard this fact and build HOBIT purely from scratch. In particular, HOBIT explicitly uses as inputs the output of the SecVtx algorithm set to its loose operating point [4] , as well as inputs to the earlier Bness [6] and Roma [7, 8] multivariate taggers. Consequently, descriptions of these taggers are given as follows:

2.1 SecVtx

SecVtx is a secondary vertex tagger. It has traditionally been the most commonly used b tagger at CDF. Using only tracks which are significantly displaced from the primary vertex, pass certain quality requirements, and are within a distance of $\Delta R < 0.4$ of a

jet's axis (with $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$), an iterative method is used to fit a secondary vertex within the jet, with the χ^2 of the fit used to guide the process. Given the relatively long lifetime of the B hadron, the significance of the two-dimensional decay length $L_{\rm xy}$ in the r-ϕ plane is used to select b jet candidates. The algorithm can be performed with different sets of track requirements and threshold values. In practice, three operating points are used, referred to as "loose", "tight", and "ultra tight". The loose operating point has not only been used standalone, but also as an input to the Roma tagger; due to its relatively higher efficiencies compared to the other two SecVtx operating points, it is also used as an input into HOBIT.

2.2 Soft Lepton Taggers

Soft lepton taggers [5] (SLT) take a different approach to b tagging. Rather than focusing on tracks within a jet, they identify semileptonic decays by looking for a lepton inside a jet's cone. The branching ratio of approximately 10% per lepton makes this method useful; however, if used alone, this class of tagger is not competitive with SecVtx or the taggers which will be described below. Because a soft lepton tagger does not rely on the presence of displaced tracks or vertices, it has a chance to identify b jets that the other methods cannot. In practice, at CDF only the soft muon tagger is used since high-purity electron or tau identification within jets is very difficult. The soft muon tagger is used to identify muons within a jet, whose multiplicities and momenta transverse to the jet axis were used as inputs to Roma, and are used now as inputs to HOBIT.

2.3 The Roma Tagger

Neural networks (NNs) can use as many flavor-discriminating observables as is computationally feasible; hence the efficiency of NN taggers is often equal to or greater than that of conventional taggers for a given purity. One such NN, the "Roma tagger", has been used at CDF in light Higgs searches [7, 8] . While the SecVtx tagger attempts to find exactly one displaced vertex in a jet, the Roma tagger uses a vertexing algorithm that can find multiple vertices, as may be the case when multiple hadrons decay within the same jet cone (for example, in a $B \to D$ decay). Three types of NNs are used: one to distinguish vertices which appear to come from heavy-flavor hadron decay from those which do not, another to distinguish unvertexed tracks which appear to come from a heavy-flavor hadron decay from those which do not, and a third that takes as inputs the first two NN outputs along with other flavor discriminating information, including loose SecVtx tag status, number of SLT-identified muons, and vertex displacement and mass information. The performance of the Roma tagger is not only superior to SecVtx at equivalent purities (see Fig. 5 for more), but it allows for an "ultra loose" operating point yielding greater efficiency, particularly useful in light Higgs searches. A majority of the inputs into the Roma tagger are employed as inputs into the HOBIT tagger, allowing HOBIT to take advantage of the same extensive vertex information of which

Roma takes advantage.

2.4 The Bness Tagger

While the Roma tagger employs a great deal of information on the vertices it finds in a jet, in the event that it is unable to fit any vertices, it is unable to distinguish b jets from light jets. However, a significant proportion of b jets (on the order of 20% in Higgs candidate events) do not contain a sufficient number of well-reconstructed tracks to allow for a vertex fit in Roma. The Bness tagger was developed not only to extract vertex information from a jet, but also to determine whether a jet is b-like based solely on the properties of its individual tracks (the Roma tagger can only examine individual tracks based on their proximity to a fitted vertex). This is done through the use of a NN which is applied to all tracks passing very loose requirements, and which takes positional (impact parameter, e.g.) and kinematic (p_T, e.g.) information on a track to determine whether it appears to have come from the decay of a B hadron displaced from the primary vertex. The Bness tagger therefore is able to extract information from all but a few percent of B jets, and can achieve a very high efficiency for a reasonable level of purity. This robust property of the tagger makes it useful for analyses with very little signal where efficiency is of the essence, as is the case with light Higgs analyses or even searches for hadronic decays of heavy gauge bosons; see [9] for more details. A very similar track-by-track NN to that employed by the Bness tagger is used to evaluate tracks in HOBIT; this will be described later in the article.

3 The CDF Detector

The CDF II detector is described in detail elsewhere [10] . The detector is cylindrically symmetric around the proton beam line[10] with tracking systems that sit within a superconducting solenoid which produces a 1.4 T magnetic field aligned coaxially with the $p\bar{p}$ beams. The Central Outer Tracker (COT) is a 3.1 m long open cell drift chamber which performs up to 96 track measurements in the region between 0.40 and 1.37 m from the beam axis, providing coverage in the pseudorapdity region $|\eta| \leq 1.0$ [11] . Sense wires are arranged in eight alternating axial and $\pm 2^\circ$ stereo "superlayers" with 12 wires each. The position resolution of a single drift time measurement is about 140 μm.

Charged-particle trajectories are found first as a series of approximate line segments in the individual axial superlayers. Two complementary algorithms associate segments lying on a common circle, and the results are merged to form a final set of axial tracks.

[10]The proton beam direction is defined as the positive z direction. The polar angle, θ, is measured from the origin of the coordinate system at the center of the detector with respect to the z axis, and ϕ is the azimuthal angle. Pseudorapidity, transverse energy, and transverse momentum are defined as $\eta = -\ln\tan(\theta/2)$, $E_T = E\sin\theta$, and $p_T = p\sin\theta$, respectively. The rectangular coordinates x and y point radially outward and vertically upward from the Tevatron ring, respectively.

Track segments in stereo superlayers are associated with the axial track segments to reconstruct tracks in three dimensions.

The efficiency for finding isolated high-momentum tracks is measured using electrons from $W^{\pm} \to e^{\pm}\nu$ decays identified in the central region $|\eta| \leq 1.1$ using only calorimetric information from the electron shower and the missing transverse energy. In these events, the efficiency for finding the electron track is $99.93^{+0.07}_{-0.35}\%$, and this is typical for isolated high-momentum tracks from either electronic or muonic W decays contained in the COT. The transverse momentum resolution of high-momentum tracks is $\delta p_T/p_T^2 \approx 0.1\%\,(\mathrm{GeV}/c)^{-1}$. Their track position resolution in the direction along the beam line at the origin is $\delta z \approx 0.5\,\mathrm{cm}$, and the resolution on the track impact parameter, the distance from the beam line to the track's closest approach in the transverse plane, is $\delta d_0 \approx 350\ \mu\mathrm{m}$.

A five layer double-sided silicon microstrip detector (SVX) covers the region between 2.5 to 11 cm from the beam axis. Three separate SVX barrel modules along the beam line cover a length of 96 cm, approximately 90% of the luminous beam interaction region. Three of the five layers combine an r-ϕ measurement on one side and a 90° stereo measurement on the other, and the remaining two layers combine an r-ϕ measurement with small angle stereo at $\pm 1.2°$. The typical silicon hit resolution is 11 μm. Additional Intermediate Silicon Layers (ISL) at radii between 19 and 30 cm from the beam line in the central region link tracks in the COT to hits in the SVX.

Silicon hit information is added to COT tracks using a progressive "outside-in" tracking algorithm in which COT tracks are extrapolated into the silicon detector, associated silicon hits are found, and the track is refit with the added information of the silicon measurements. The initial track parameters provide a width for a search road in a given layer. Then, for each candidate hit in that layer, the track is refit and used to define the search road into the next layer. This stepwise addition of precision SVX information at each layer progressively reduces the size of the search road, while also accounting for the additional uncertainty due to multiple scattering in each layer. The search uses the two best candidate hits in each layer to generate a small tree of final track candidates, from which the tracks with the best χ^2 are selected. The efficiency for associating at least three silicon hits with an isolated COT track is $91 \pm 1\%$. The extrapolated impact parameter resolution for high-momentum outside-in tracks is much smaller than for COT-only tracks: 30 μm, including the uncertainty in the beam position.

Outside the tracking systems and the solenoid, segmented calorimeters with projective geometry are used to reconstruct electromagnetic (EM) showers and jets. The EM and hadronic calorimeters are lead-scintillator and iron-scintillator sampling devices, respectively. The central and plug calorimeters are segmented into towers, each covering a small range of pseudorapidity and azimuth, and in full cover the entire 2π in azimuth and the pseudorapidity regions of $|\eta|<1.1$ and $1.1<|\eta|<3.6$ respectively. The transverse energy E_T, where the polar angle is calculated using the measured z position of the event vertex, is measured in each calorimeter tower. Proportional and scintillating strip detectors measure the transverse profile of EM showers at a depth

corresponding to the shower maximum.

High-momentum jets, photons, and electrons leave isolated energy deposits in contiguous groups of calorimeter towers which can be summed together into an energy cluster. Electrons are identified in the central EM calorimeter as isolated, mostly electromagnetic clusters that also match with a track in the pseudorapidity range $|\eta| < 1.1$. The electron transverse energy is reconstructed from the electromagnetic cluster with precision $\sigma(E_T)/E_T = 13.5\%/\sqrt{E_T(\text{GeV})} \oplus 2\%$, where the $\oplus$ symbol denotes addition in quadrature. Jets are identified as a group of electromagnetic and hadronic calorimeter clusters using the JETCLU algorithm [12] with a cone size of 0.4. Jet energies are corrected for the calorimeter non-linearity, losses in the gaps betwen towers, multiple primary interactions, the underlying event, and out-of-cone losses [13] . The jet energy resolution is approximately $\sigma_{E_T} = 1.0 \text{ GeV} + 0.1 \times E_T$

Directly outside of the calorimeter, four-layer stacks of planar drift chambers detect muons with $p_T > 1.4$ GeV/c that traverse the five absorption lengths of the calorimeter. Farther out, behind an additional 60 cm of steel, four layers of drift chambers detect muons with $p_T > 2.0$ GeV/c. The two systems both cover a region of $|\eta| \leq 0.6$, though they have different structure and their geometrical coverages do not overlap exactly. Muons in the region between $0.6 \leq |\eta| \leq 1.0$ pass through at least four drift layers lying in a conic section outside of the central calorimeter. Muons are identified as isolated tracks in the COT that extrapolate to track segments in one of the four-layer stacks.

4 The HOBIT Tagger

The HOBIT tagger contains all the same advantages of other multivariate taggers such as Roma [7, 8] and Bness [6] , most noteably a near-maximal use of the information available in a b jet and a tunable purity-efficiency curve. Unlike other taggers, however, HOBIT was constructed so as to be optimized for light Higgs searches. Its training was performed using b jets in Pythia [14] 120 GeV Higgs Monte Carlo (MC) and light jets from Pythia W+jets MC; charm jets were ignored based on preliminary studies which indicated a relative insensitivity of light Higgs searches to charm jet contamination. Here, "b jet" denotes a jet with a B hadron within a cone of ΔR <0.4 of its axis, while a charm jet contains a charm hadron (but no B hadrons) within this cone and a light jet contains neither B hadrons nor charm hadrons within its cone. Jets were required to have an $E_T > 15$ GeV, $|\eta| < 2$, and at least one Bness-selected track.

HOBIT is constructed as a feed-forward multilayer perceptron neural network implemented using the TMVA package for Root [15] . It consists of two hidden layers of N and $N + 1$ nodes, $N = 25$ being the number of inputs to the tagger, and a tanh activation function; 500 cycles were used in the training. Broadly speaking, the inputs to the tagger can be considered to be either Roma inputs or Bness inputs. Most of the Roma inputs are used in HOBIT; these include properties of the fitted vertex found to be the most heavy-flavor-like (its displacement, invariant track mass, pseudo-cτ), as well as the number of SLT-tagged muons and the jet's loose SecVtx tagger status.

In addition to these inputs, the ten tracks in the jet cone found to have the highest track bness values have those values input into HOBIT. In the event that there are only N tracks in a jet, where N<10, the N+1-th highest track bness, etc., is set to -1. Note that the selection on tracks used for the Roma and Bness inputs differ; the same track selection as was originally used for Roma is used for the Roma inputs here (tracks must have a $p_T > 1$ GeV and be within $\Delta R < 0.4$ of the jet axis) while for the track-by-track Bness inputs a very loose selection requiring a $p_T > 0.5$ GeV and a distance of $\Delta R < 0.7$ of the jet axis is used. Other selection cuts were tried, but none resulted in an improvement in the performance of HOBIT.

The inputs to the track-by-track Bness NN are the same here as were used in the original Bness tagger; however, note that the NN was retrained both due to the looser cone requirement (the original Bness tagger required $\Delta R < 0.4$) and due to the desire to optimize it for light Higgs searches. Hence, the track Bness NN was trained off of the same MC as was used to train the overall HOBIT tagger, light Higgs for b jets, and W+jets for light jets. An additional requirement on the b jet tracks used in training is that they be within $\Delta R < 0.05$ of the actual charged particles resulting from a B hadron decay in the MC. The same basic framework as was used in the training of HOBIT itself (training cycles, inner layer structure, etc.) was used in the training of the track-by-track Bness NN.

As one would expect, the inputs take advantage of the fact that tracks from B hadron decays are displaced from the primary vertex (the impact parameter and ΔZ between a track and the primary vertex, as well as their significances). Furthermore, kinematic inputs such as the p_T, rapidity, and p_{perp} with respect to the jet axis take advantage of the greater collimation of B tracks due to the large boost of the hadron. Finally, the jet E_T itself is an input to the track-by-track NN; this is because previously mentioned inputs are correlated with jet E_T and hence jet E_T will provide additional information for track discrimination. Note that tracks from light jets are weighted in training so as to have the same parent jet E_T distribution as tracks from b jets; this is done so as to avoid kinematic biasing in the NN. Distributions of the track-by-track inputs are shown in Fig. 1; not shown are the jet E_T distributions which are identical by construction. Additional separation power for HOBIT was found with the addition of the number of tracks which pass the selection cuts required for track-by-track Bnesses; hence this number is added as an input to HOBIT as well.

Roma inputs used in HOBIT consist of observables built off of tracks and vertices found to be HF-like using the relevant NNs. These include:

- The invariant mass, pseudo-cτ, 3-d displacement and 3-d displacement significance of the most HF-like vertex
- The number of tracks both in HF-like vertices and standalone HF-like tracks, as well as their invariant mass, and the ratio of the p_T's of these tracks to the p_T's of all tracks in the jet
- The loose SecVtx tag status, as well as the mass of the tracks used in the loose

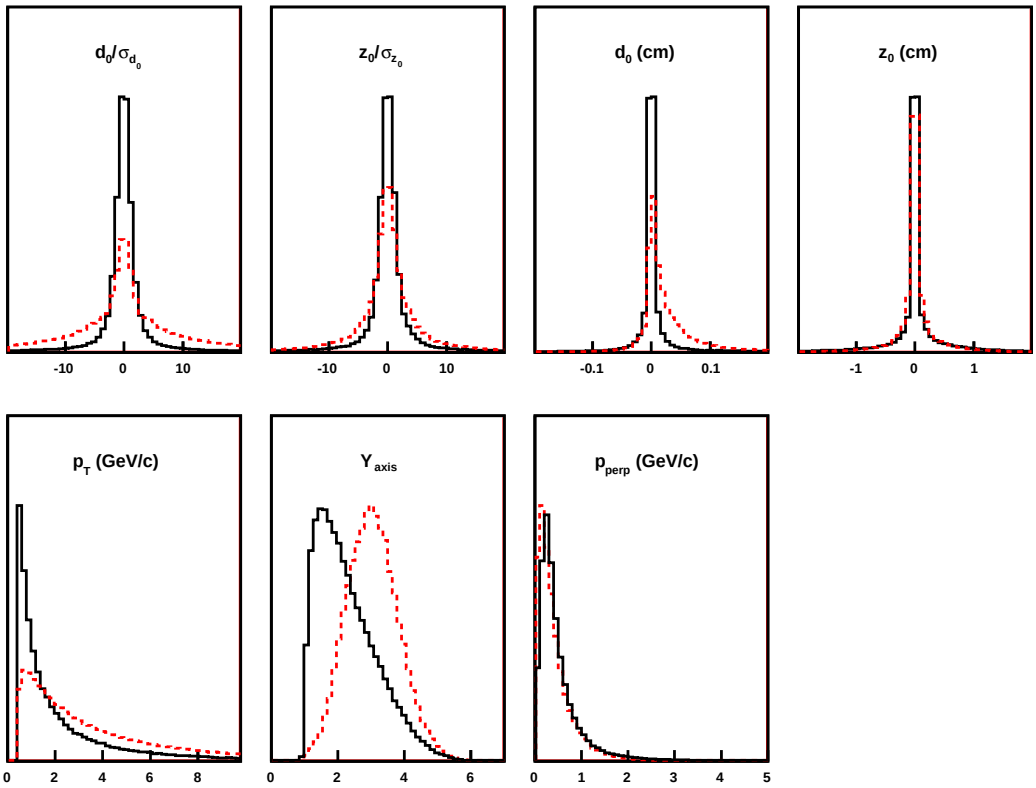

Figure 1: Inputs to track-by-track Bness. The solid histogram is for tracks in light quark jets and the dashed (colored) histogram is for tracks in b jets. Not shown is the jet E_T, identical between the two distributions by construction.

SecVtx vertex fit

As mentioned, one potential weakness of the Roma tagger is its requirement that a jet have at least one HF-like vertex; jets without such a vertex are ignored. This requirement of "Roma taggability" can be a liability when very high b jet tagging efficiency is sought; in the MC sample used to train the HOBIT tagger, while 31% of light jets fail to be Roma taggable, 23% of b jets also fail, a disadvantage when one wants to have a loose, highly efficient tagger and is willing to allow some light jets to be passed. This weakness is compensated for via the track-by-track Bnesses; while jets in HOBIT are required to have at least one track with an evaluated Bness, only 3.0% of b jets and 2.1% of light jets in the MC fail this requirement, indicating a very flexible taggability requirement. A final input to HOBIT is the E_T of the jet itself. As other inputs are correlated with E_T, the E_T should provide additional useful information to HOBIT. As is the case with the track-by-track Bness, kinematic biasing of HOBIT is prevented by weighting the light jets so as to have the same E_T distribution as the b jets. Distributions of the inputs to HOBIT are shown in Fig. 2. The output

HOBIT distributions for b jets and light jets from the same type of MC as was used to train the discriminator are shown in Fig. 3. In Fig. 4, the b-jet efficiencies and the mistag rates as a function of jet E_T and η are shown for two HOBIT operating points – a cut at 0.72 ("loose") and a cut at 0.98 ("tight"). At higher η, where tracking coverage is more sparse and less information is available, the b-tagging efficiency drops, as would be expected; interestingly, the mistag rate increases in the case of the loose tag and drops in the case of the tight tag, demonstrating the greater damage incorrectly identified tracks can cause when tagging is loose. As a function of jet E_T, in general the efficiency increases with increasing jet E_T due to the greater displacement of the B hadron; similarly the light jet efficiency increases, perhaps in part due to the greater error on the momentum measurements of the tracks and perhaps also due to actual long lived particles in light jets (such as K_s and Λ) having a greater displacement and thus being misidentified as B hadrons.

In order to evaluate the performance of HOBIT, we compare its purity-efficiency curve to the curves of the Bness and Roma taggers; additionally, we look at the purity/efficiency performance of SecVtx at both its tight and loose operating points. Here, purity refers to the fraction of light jets in W+jets MC which don't get tagged as b jets, and efficiency refers to the fraction of b jets in light Higgs MC which do get tagged. The jets in both the numerator and denominator when evaluating tag efficiencies are required to have an $E_T > 15$ GeV, with $|\eta| < 2$; these are the same E_T and η requirements as were placed on the jets in the training of HOBIT. Results are in Fig. 5; note that for a given purity level, HOBIT results in a roughly 10% absolute efficiency improvement over the Bness and Roma taggers, and a roughly 15% improvement over the SecVtx taggers.

One interesting question to investigate is how much of the improvement in HOBIT is due to the Higgs optimization. To study this, we compared the purity-efficiency curves of the original Bness and Roma taggers with NN taggers we trained off of W+jets and light Higgs MC which take the same inputs as Bness and Roma. The results can be seen in Figs. 6 and 7. Note that in the case of the Roma comparison, not only is our retrained Roma tagger compared with the original Roma result, but also with Roma's b vs. light jet separator. This is because the architecture of Roma consisted of three different NN separators (b vs. light, b vs. charm, light vs. charm) which fed into the final Roma NN. As we retrained off of light and b jets, the comparison of our Higgs-optimized version of Roma with their b vs. light separator makes for a more fair comparison. In both the Bness and Roma cases, the improvement in efficiency appears to be around 2%, absolute.

5 Efficiency and Mistag Scale Factors

In order to be used in a physics analysis, the performance of the HOBIT b-tagger must be calibrated. Historically, the MC modeling of the b-tag efficiencies and mistag rates has not been sufficient to use the predictions of the MC directly in physics analyses.

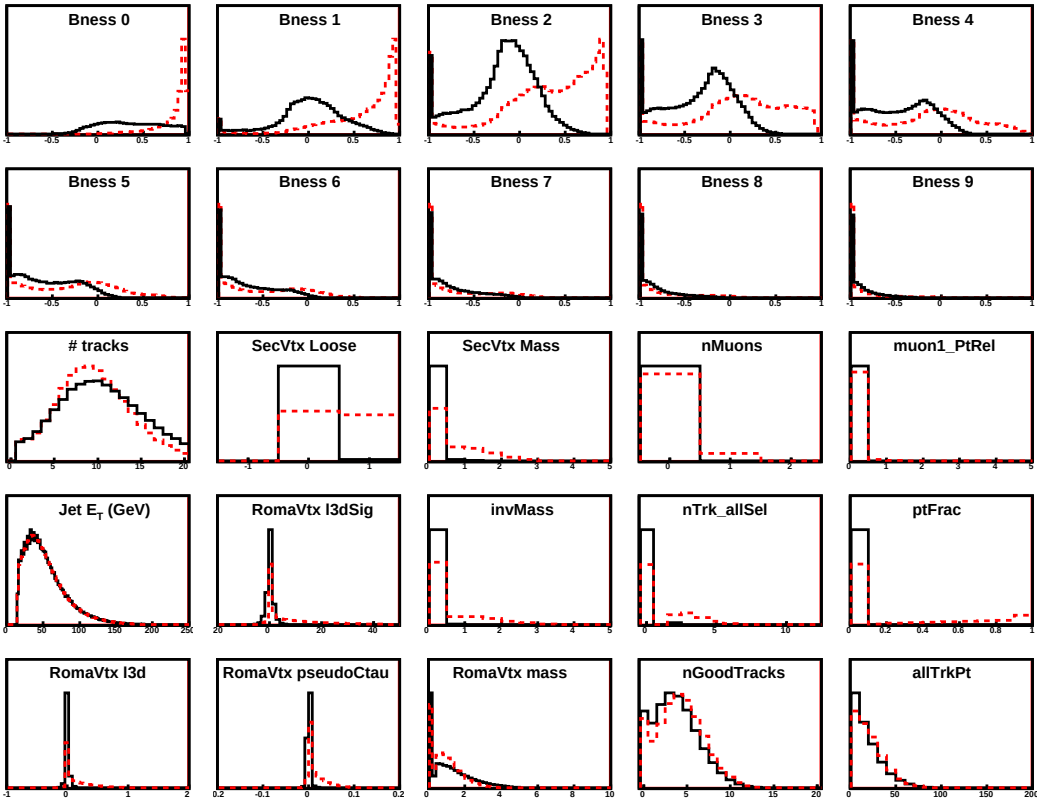

Figure 2: Inputs to HOBIT. The solid histogram is for light quark jets and the dashed (colored) histogram is for b jets. Left to right, top to bottom: the ten highest Bness tracks, the # of Bness-selected tracks, the loose SecVtx tag status and the mass of its fitted vertex, the # of SLT-tagged muons and the momentum of the most SLT-favored muon transverse to the jet axis, jet E_T, the 3-d displacement significance of the most heavy-flavor-like vertex in Roma, the invariant mass, raw count, and fraction of total track P_T of heavy-flavor-like tracks, the 3-d displacement, pseudo-cτ and invariant mass of the most heavy-flavor-like vertex, the # of Roma-selected tracks and their total P_T.

Instead, techniques have been developed to measure the b-tagging efficiency using CDF data, and also to constrain the mistag rate. Examples are using jets containing electrons (hence, HF-enriched) [16] for checking the b-tagging efficiency of the SecVtx algorithm, and by using the rate at which jets have a secondary vertex reconstructed behind the primary vertex ("negative tags") in order to estimate mistags [17]. For the tight SecVtx tagger, the b-tag efficiency is found to be well predicted by the MC up to a scale factor, which has a value of 0.96 ± 0.05 for the full CDF dataset. We seek here to provide the same level of detail and systematic control over the tag performance of HOBIT, for each operating point.

An important difference between SecVtx and HOBIT is the absence of negative

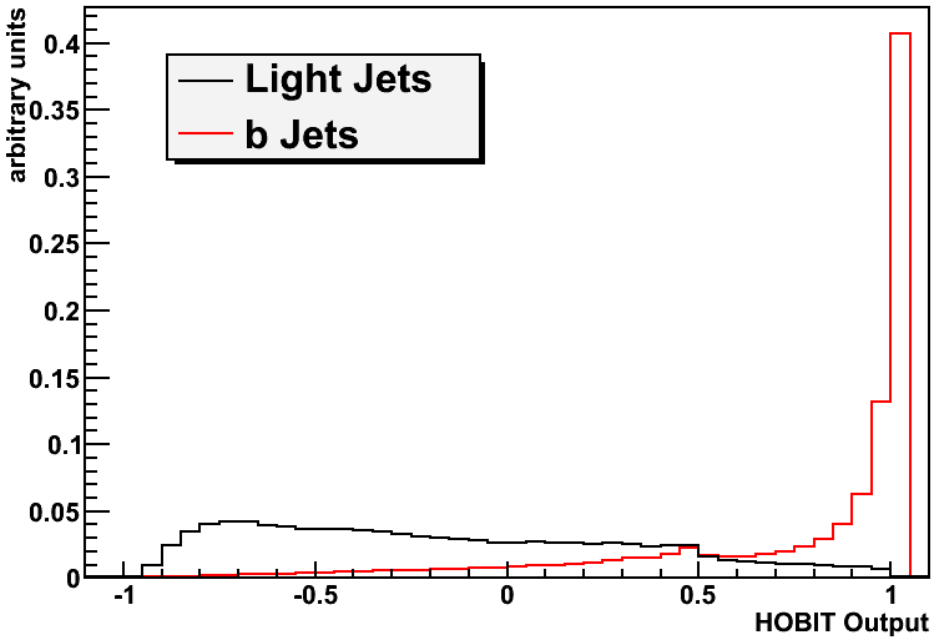

Figure 3: HOBIT outputs. The black histogram is for light quark jets and the colored histogram is for b jets.

tags in HOBIT, meaning the SecVtx mistag calculation technique is inapplicable. Instead, two new techniques were developed and are described below for calibrating b-tag scale factors and providing mistag rates: the $t\bar{t}$ cross section method, and the electron conversion method.

5.1 Scale factors using the $t\bar{t}$ cross section method

The $t\bar{t}$ cross section method seeks to calibrate scale factors for the b-tagging efficiency and the mistag rate relative to MC predictions. Tag-rate matrices similar to the SecVtx mistag matrix are constructed separately for b, c, and light jets. These matrices are filled in with MC HOBIT tag rates in $t\bar{t}$ MC samples for the b and c jets, and for W+jets samples for the light-flavored jets. A separate matrix is made for each flavor of jet for each HOBIT operating point. The same variables are used in the same binning as the SecVtx mistag matrix, except the $\sum E_T$ variable is now omitted. The remaining variables parameterizing the matrices are jet E_T, jet η, the number of tracks in the jet, the number of primary vertices, and the z location of the primary vertex that the jet is assigned to have come from.

The tag rates are expected to be different in data and MC by a scale factor, which is calibrated by comparing the predicted and observed event counts in single- and double-

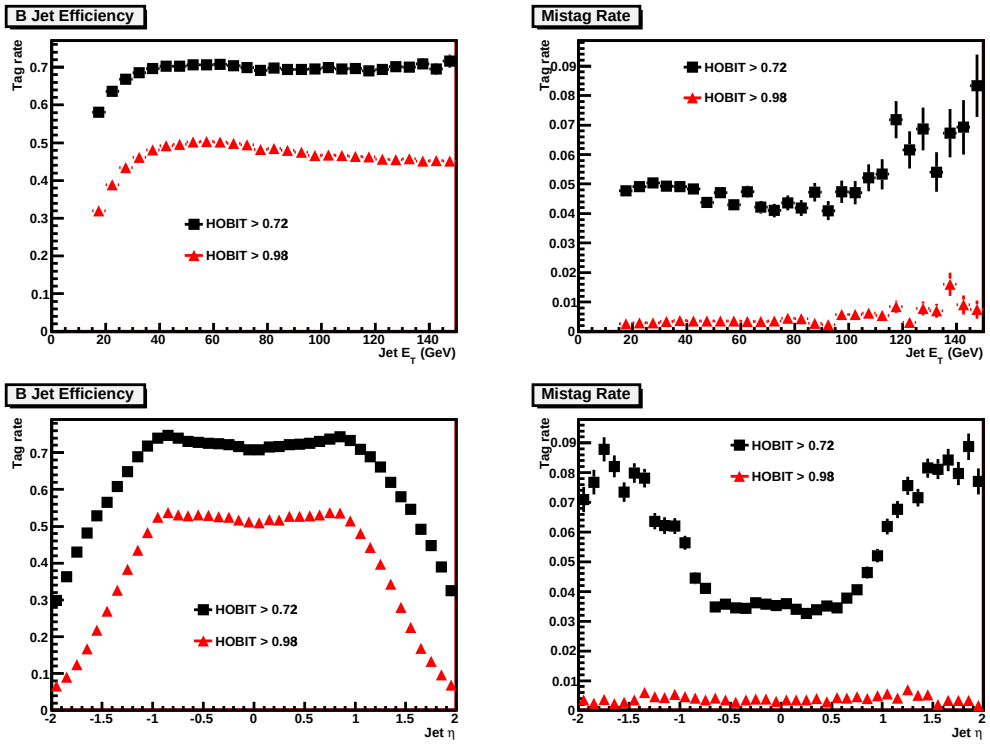

Figure 4: b jet and light jet efficiencies as a function of η and E_T. The black triangles are for the looser operating point and the colored triangles are for the tighter operating point.

HOBIT tagged W+jets events. The method is based on the simultaneous measurement of the SecVtx b-tag scale factor and the $t\bar{t}$ cross section [18]. In this case, the single- and double-tagged W+3 or more jet samples provide two constraints which allow the measurement of two unknowns. A two-dimensional fit is performed to maximize the likelihood of observing the data counts as functions of the SecVtx b-tag scale factor and the $t\bar{t}$ cross section.

This method has been repurposed to measure the HOBIT b-tag scale factor and the HOBIT mistag scale factor, where the $t\bar{t}$ cross section is now an input assumption. In [18], the W+3 or more jet sample is used with standard W+jets event selection, but with an additional cut of $H_T > 250$ GeV, in order to purify the sample in $t\bar{t}$ events. This sample has insufficient mistags in it to calibrate the mistag scale factor adequately, and so the $W + 1$ jet sample has been added, without the H_T cut. This sample is almost pure W+light flavor (LF) events, but after b-tagging, consists of comparably-sized $Wb\bar{b}$, $Wc\bar{c}$, Wcj, and mistagged W+LF events. The background predictions [4] scale the total W+jets rate and subtract off the non-W+jets components, but the prediction of the W+HF components relies on the heavy-flavor K-factor. We find that the $W+1$-jet data provides a rather independent handle on the mistag scale factor; the

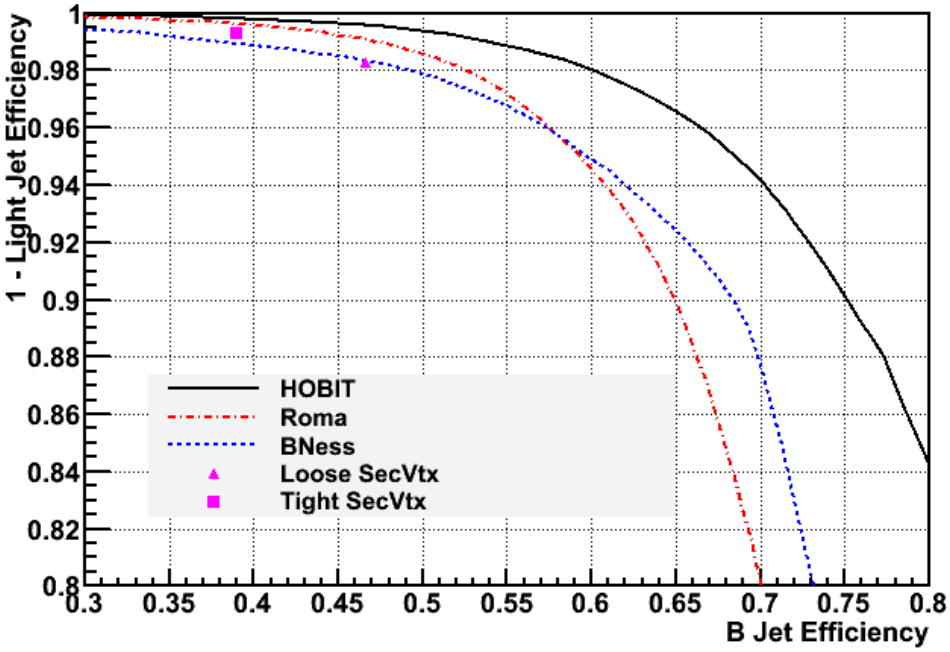

Figure 5: A comparison of the purity-efficiency tradeoffs for HOBIT vs. Roma, Bness, and SecVtx loose and tight. A significant improvement over prior multivariate taggers is seen.

b-tag scale factor is constrained well by the events with three or more jets. However, the dependence on the heavy-flavor K-factor introduces a systematic uncertainty that strongly affects the mistag scale factor. For low values of the HOBIT cut, the mistag rate is relatively high, and the contribution to the tagged W+1-jet sample from W+HF events is lower, and thus the systematic uncertainty on the mistag scale factor due to the uncertainty on the HF K-factor is less at low HOBIT cut values than at high HOBIT cut values.

The likelihood of observing the data given the b-tag scale factor and the mistag scale factor for fixed values of the HF K-factor, the $t\bar{t}$ cross section, and the HOBIT cut is then maximized over the two scale factors. The dependence on the HF K-factor and the $t\bar{t}$ cross section are then taken as sources of systematic uncertainty. We assume $\sigma_{t\bar{t}} = 7.04 \pm 0.704$ pb [19], and take the HF K-factor to be 1.4 ± 0.4. The fitted b-tag and mistag scale factors are shown in Figures 8 and 9, respectively, as functions of the HOBIT cut. A line is fit to the b-tag scale factor as a function of the HOBIT cut, and a parabola is fit to the mistag scale factor. The variation due to $\sigma_{t\bar{t}}$ is also shown, and symmetrized about the larger variation.

The determination of the b-tag and mistag scale factors are subject to the same sources of systematic uncertainty as a measurement of $\sigma_{t\bar{t}}$; see Ref. [20]. Namely, the

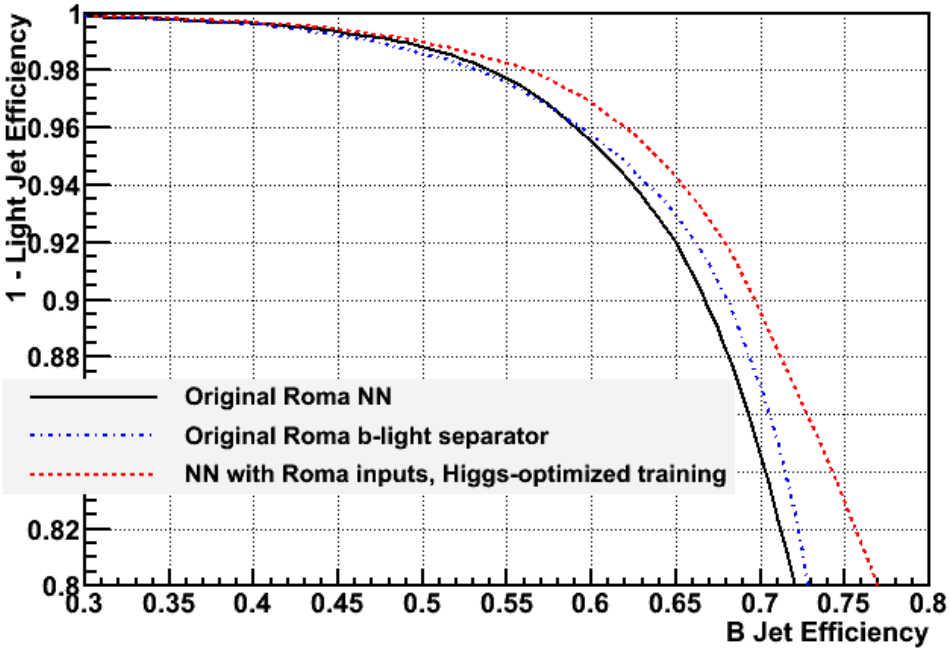

Figure 6: A comparison of the purity-efficiency tradeoffs for the original Roma tagger (as well as its b-light separator; see text for more) and our version of the Higgs-optimized Roma tagger.

$t\bar{t}$ acceptance depends on ISR+FSR (initial-state radiation and final-state radiation), PDF (parton distribution functions), JES (jet energy scale), and trigger and lepton identification. The b-tag efficiency is being calibrated here, and the luminosity uncertainty, nearly absent in Ref. [20], returns here. All sources of systematic uncertainty have the same impact on this method as changing the assumed $\sigma_{t\bar{t}}$. Thus the dependence of the measured b-tag and mistag scale factors on the assumed value of $\sigma_{t\bar{t}}$ provides the mechanism by which the other sources of systematic uncertainty can be evaluated.

For the loose (0.72) and tight (0.98) HOBIT operating points, this method yields SFs of 0.997 ± 0.037 and 0.917 ± 0.069, respectively. A complete table of systematics for the SF is shown in Table 1, and for the mistag matrix in Table 2. Validation plots comparing properties of the highest E_T jet in $t\bar{t}$ candidate events for MC vs. data are shown in Figs. 10, 11, 12, 13, and 14.

5.2 Scale factors using the electron conversion method

Another method of calculating the SF for the HOBIT tagger involves a modification of the traditional SecVtx SF algorithm [16] , which, unlike the SecVtx technique, doesn't

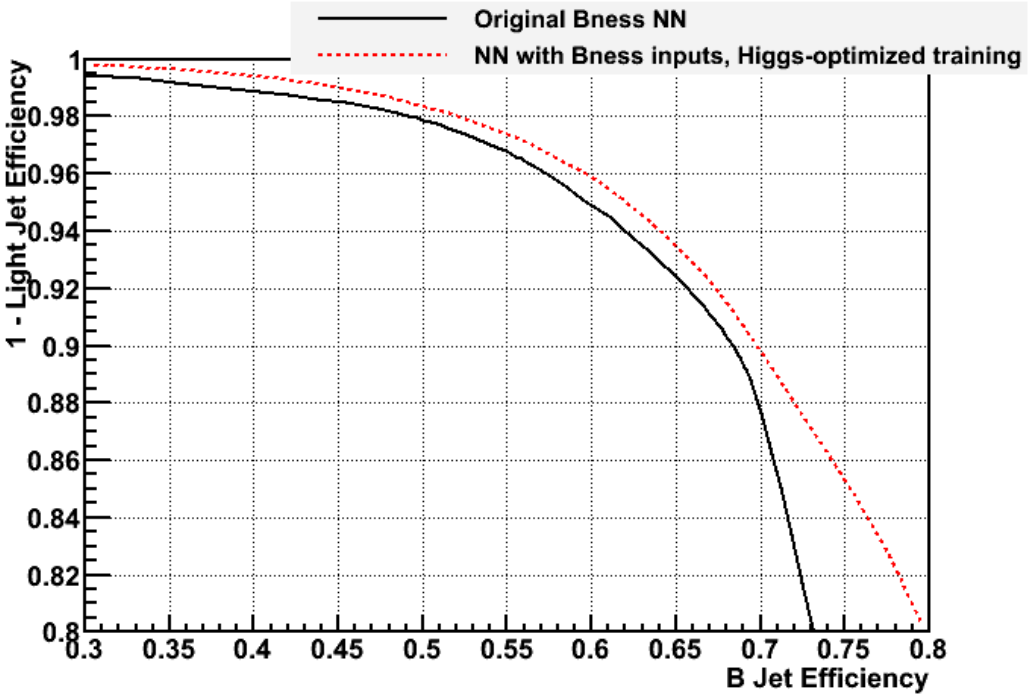

Figure 7: A comparison of the purity-efficiency tradeoffs for the original Bness tagger and our version of the Higgs-optimized Bness tagger.

require the concept of a "negative tag". However, like the SecVtx technique, this method involves taking advantage of the differing levels of heavy-flavor enhancement among jets containing electrons depending on whether the electron is identified as coming from a conversion or not.

The event sample consists of back-to-back dijet events where one jet, the "ejet" (the electron jet), contains an electron, and its opposite jet is referred to as the "ajet" (the away jet). As this method relies heavily on algebra, some notation needs to be established. We can label each jet originating either from the heavy-flavor quark (B) decay or the light-flavor (Q) and categorize events as N_{XY} where the ejet has flavor X and the ajet has flavor Y. Then the total number of selected events (N^e) is simply

$$N_{BB} + N_{BQ} + N_{QB} + N_{QQ} = N^e$$

and the heavy-flavor fraction of the ejets is

$$F_B = (N_{BB} + N_{BQ})/N^e.$$

Applying a b-tag on the ejet with a tagging efficiency (ϵ^e) and a mistag rate (ϵ_{mis}), the number of b-tagged ejets (N^e_+) is

$$\epsilon^e \cdot (N_{BB} + N_{BQ}) + \epsilon^e_{mis} \cdot (N_{QB} + N_{QQ}) = N^e_+.$$

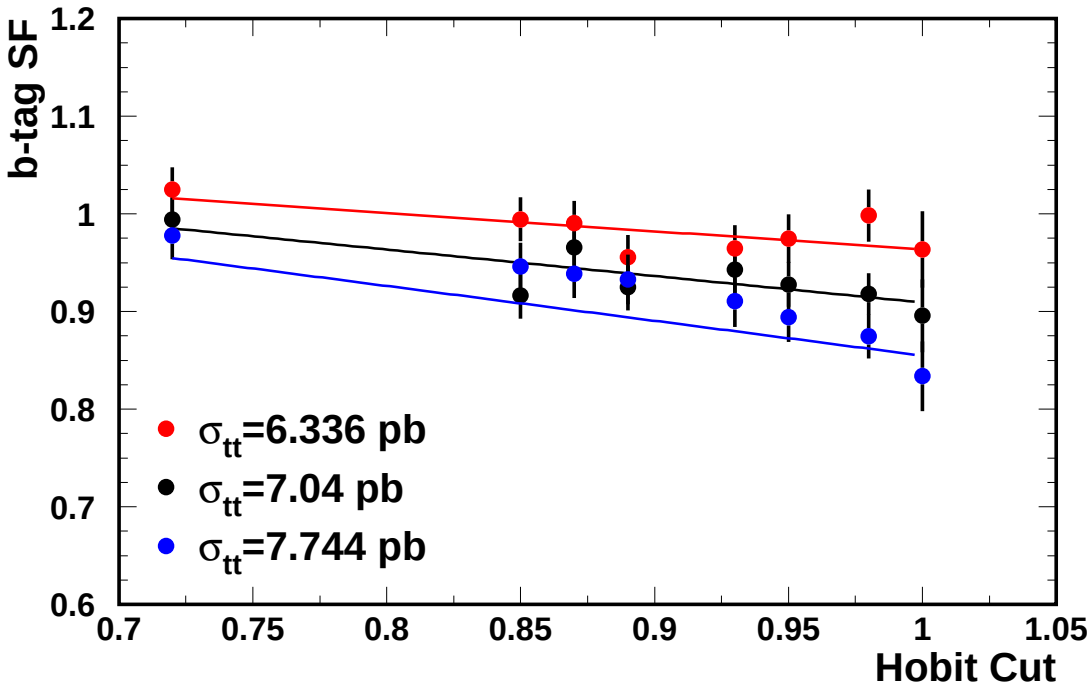

Figure 8: The measured value of the b-tag scale factor for the HOBIT tagger as a function of the HOBIT cut. Variations are shown assuming two values of the $t\bar{t}$ cross section. Straight lines are fit to the central values and the $\sigma_{t\bar{t}} = 6.336$ values, and reflected over the central line to obtain the other variation in order to be conservative.

Assuming the conversion finding efficiency is ϵ^c for the light-flavor jets and ϵ^0 for the heavy-flavor jet, we can obtain the number of ejets identified from the conversion N^{ec} as

$$\epsilon^0 \cdot (N_{BB} + N_{BQ}) + \epsilon^c \cdot f^c \cdot (N_{QB} + N_{QQ}) = N^{ec}$$

after tagging, the number of b-tagged conversion ejets (N^{ec}_+) becomes

$$k \cdot \epsilon^e \cdot \epsilon^0 \cdot (N_{BB} + N_{BQ}) + \epsilon^e_{mis} \cdot \epsilon^c \cdot f^c \cdot (N_{QB} + N_{QQ}) = N^{ec}_+,$$

where k is a ratio of b-tag efficiencies for the heavy-flavor ejet identified as conversion or not.

The previous two equations allow us to solve for ϵ_{mis} and ϵ^e:

$$\epsilon_{mis} = (N^{ec}_+ - k \cdot \epsilon^0 \cdot N^e_+)/(N^{ec} - \epsilon^0 \cdot N^e \cdot (k + (1-k) \cdot F_B))$$

and

$$\epsilon^e = (N^e_+ - \epsilon_{mis} \cdot N^e \cdot (1 - F_B))/(N^e \cdot F_B).$$

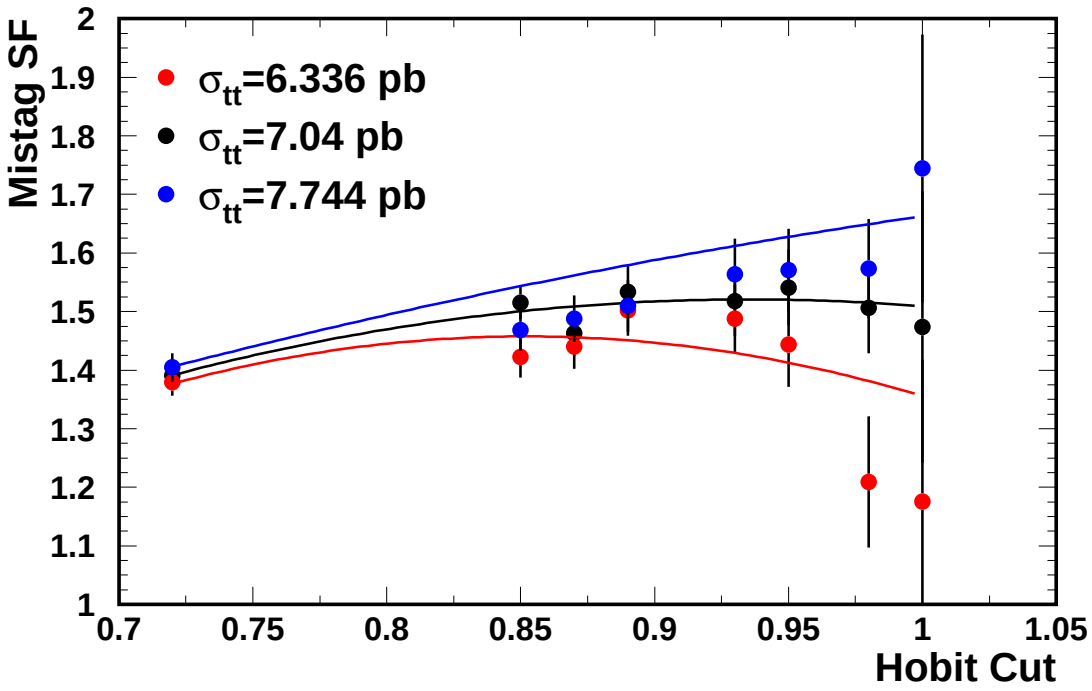

Figure 9: The measured value of the mistag scale factor for the HOBIT tagger as a function of the HOBIT cut. Variations are shown assuming two values of the $t\bar{t}$ cross section. Parabolas are fit to the central values and the $\sigma_{t\bar{t}} = 6.336$ values, and reflected over the central curve to obtain the other variation in order to be conservative.

Here, all terms that aren't the mistag and efficiency rates can be counted directly or taken from MC (k), measured (F_B) or either taken from MC or measured (ϵ^0). The resulting SFs for the loose and tight HOBIT outputs are 0.986 ± 0.066 and 0.949 ± 0.044, respectively, in good agreement with the results from the $t\bar{t}$ method. The SFs on the mistag rate calculated from the method for loose and tight are 1.28 ± 0.17 and 1.42 ± 0.89, respectively, also consistent with the results of the $t\bar{t}$ method. As a check, plots comparing ejet candidate data with MC are shown in Figs. 16 and 15, purified for heavy-flavor by requiring the away jet to be tight SecVtx tagged and the electron in the ejet to not be identified as a conversion. The fraction of HF vs. light jet MC is determined via a fit to the HOBIT distribution.

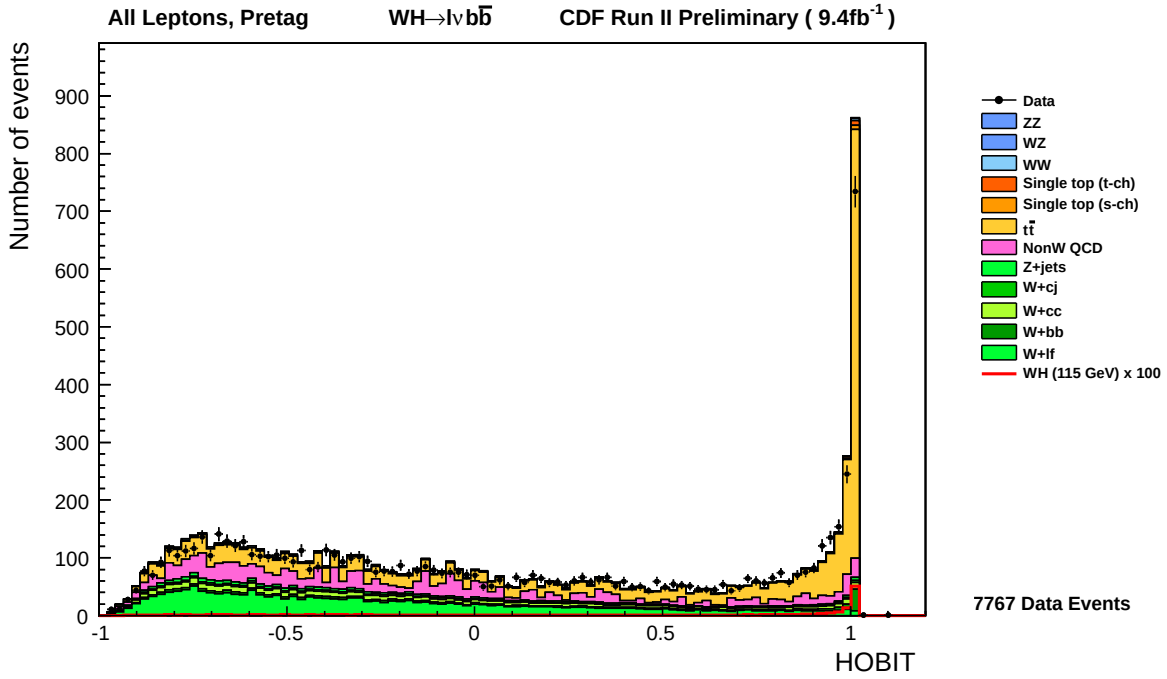

Figure 10: Data vs. MC, HOBIT output, highest E_T jet in $t\bar{t}$ candidate events

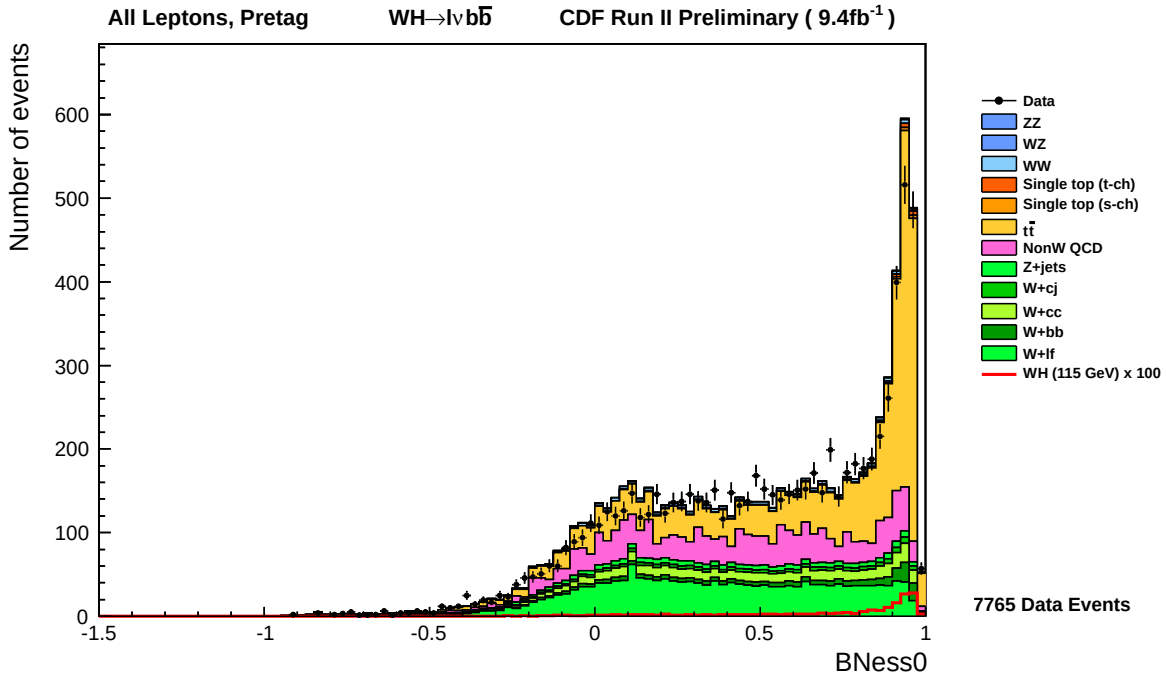

Figure 11: Data vs. MC, highest track bness, highest E_T jet in $t\bar{t}$ candidate events

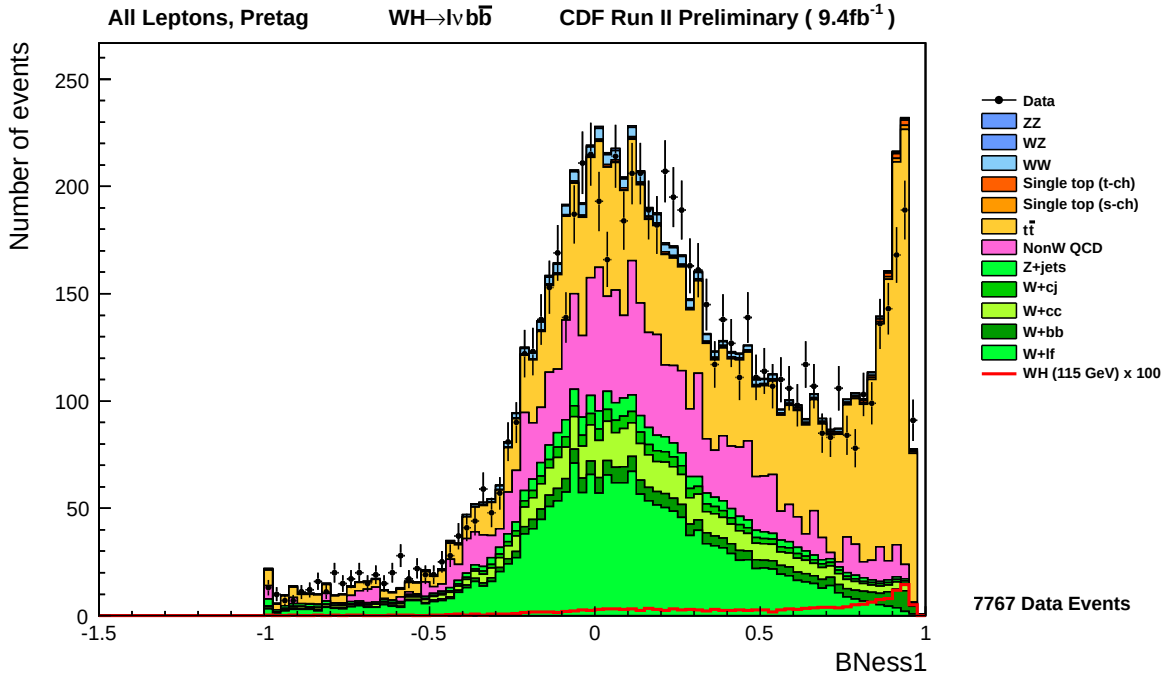

Figure 12: Data vs. MC, second highest track bness, highest E_T jet in $t\bar{t}$ candidate events

6 SF Combination

When combining the SF's calculated using the electron method and the $t\bar{t}$ method, we obtain 0.993 ± 0.032 (for HOBIT's loose operating point, 0.72) and 0.937 ± 0.037 (HOBIT's tight operating point, 0.98).

7 Conclusion

We have described a neural-network based b-identification tagger which draws on the best ideas of previous CDF taggers, has a very generous taggability requirement, and has been optimized for $H \rightarrow b\bar{b}$ searches, the primary decay channel of the light Higgs. HOBIT's scale factor has been calculated using two uncorrelated and innovative methods, both of whose answers are in good agreement; furthermore, a mistag matrix has also been calculated. In current use by light Higgs analyses at CDF, insertion of HOBIT into the analyses has resulted in a 10-20% improvement in Higgs sensitivity.

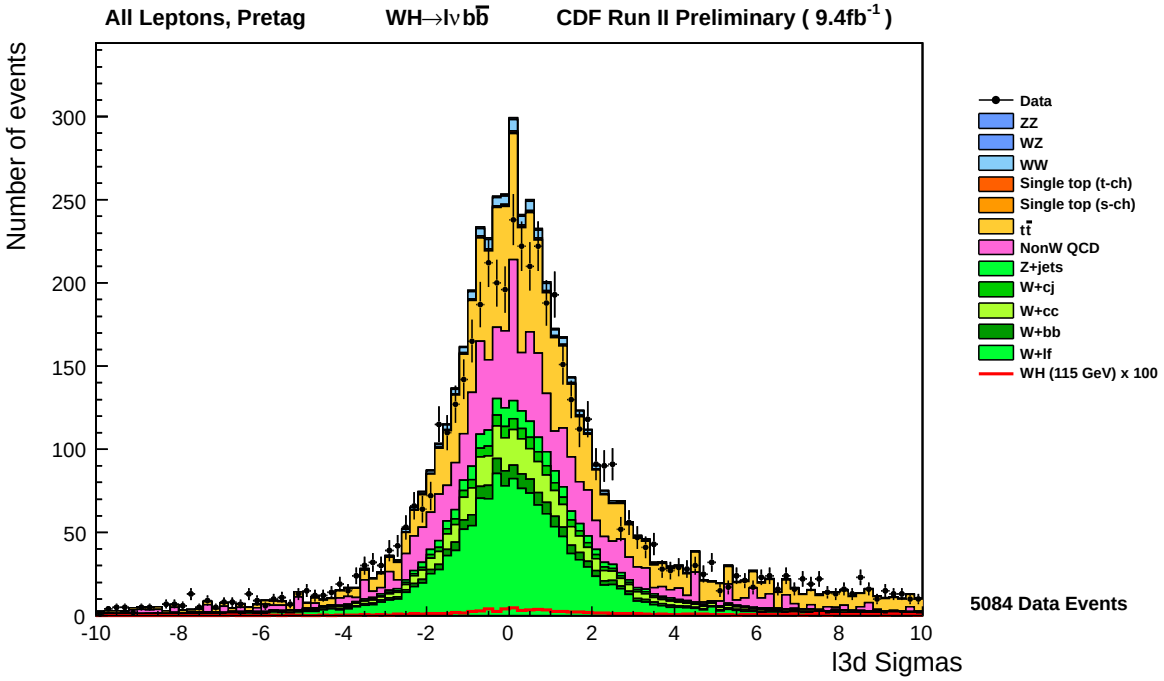

Figure 13: Data vs. MC, 3-d displacement significance of most HF-like secondary vertex, highest E_T jet in $t\bar{t}$ candidate events

References

[1] V.M. Abazov, et al., b-Jet Identification in the D0 Experiment, Nucl. Instrum. Methods A 620 (2-3) (2010) 490.

[2] CMS Collaboration, Performance of the b-jet Identification in CMS, CMS Physics Analysis Summary

[3] ATLAS Collaboration, Commissioning of the ATLAS High-Performance b-Tagging Algorithms in the 7 TeV Collision Data, ATLAS CONF Note.

[4] D. Acosta, et al., Measurement of the $t\bar{t}$ Production Cross Section in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV Using Lepton+Jets Events with Secondary Vertex b-tagging, Phys. Rev. D 71 (2005) 052003

[5] D. Acosta, et al., Measurement of the $t\bar{t}$ Production Cross Section in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV Using Lepton+Jets Events with Semileptonic B Decays to Muons, Phys. Rev. D 72 (2005) 032002

[6] J. Freeman, et al., An Artificial Neural Network Based B-Jet Identification Algorithm at the CDF Experiment, Nucl. Instrum. Meth. A, Vol. 663 (2012), pp. 27-37

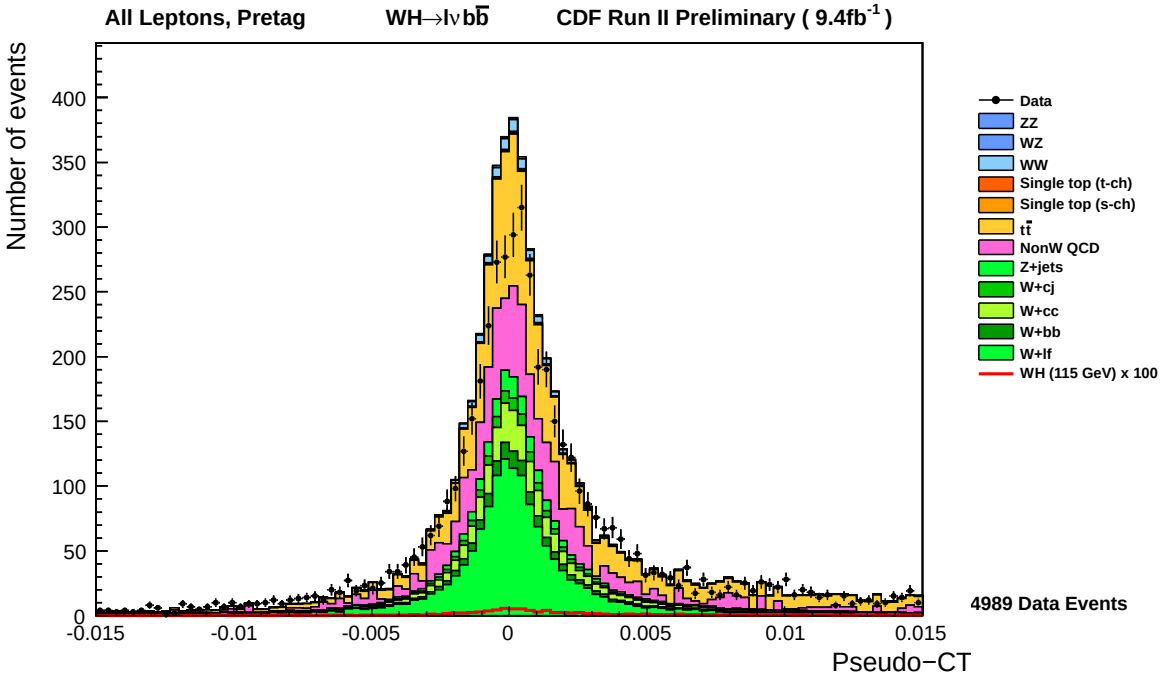

Figure 14: Data vs. MC, pseudo-cτ of most HF-like secondary vertex, highest E_T jet in $t\bar{t}$ candidate events

[7] C. Ferrazza, Identificazione di quark pesanti in getti adronici in interazioni $p\bar{p}$ con il rivelatore CDF al Tevatron, Master's thesis, Universita "La Sapienza" Roma (2006).

[8] P. Mastrandrea, Study of the heavy flavour fractions in Z+jets events from $p\bar{p}$ collisions at energy = 1.96 TeV with the CDF II detector at the Tevatron collider, FERMILAB-THESIS-2008-63.

[9] T. Aaltonen, et al., Search for $WZ + ZZ$ Production with Missing Transverse Energy+Jets with b Enhancement at $\sqrt{s}$ =1.96 TeV, Phys. Rev. D 85 (2012) 012002

[10] A. Abulencia, et al., Measurements of inclusive W and Z cross sections in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV, J. Phys. G 34 (2007) 2457

[11] T. Affolder, et al., CDF Central Outer Tracker, Nucl. Instrum. Methods A 526 (3) (2004) 249.

[12] F. Abe, et al., Topology of three-jet events in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV, Phys. Rev. D 45 (1992) 1448.

[13] A. Bhatti, et al., Determination of the jet energy scale at the Collider Detector at Fermilab, Nucl. Instrum Methods A 566 (2) (2006) 375.

Table 1: The full list of systematic uncertainties for the b-jet tagging efficiency scale factor from the $\sigma(t\bar{t})$ method measurement. This uncertainty must be combined with the electron method scale factor uncertainty and the two should be treated as uncorrelated. The uncertainties shown below are absolute shifts, and thus should be applied as $SF \pm shift$, e.g. 0.993 ± 0.022.

b-eff SF $\sigma(t\bar{t})$ method		HOBIT Operating Point	
source		Loose	Tight
$\sigma(t\bar{t})$	up	-0.011	-0.019
	down	0.011	0.019
luminosity	up	-0.004	-0.055
	down	0.007	0.012
jet energy scale	up	-0.005	-0.007
	down	0.005	0.007
generator	up	0.003	0.005
	down	-0.003	-0.005
generator	up	0.003	0.005
	down	-0.003	-0.005
ISR/FSR	up	-0.001	-0.001
	down	0.001	0.001
$t \to Wb$ branching ratio	up	-0.001	-0.001
	down	0.001	0.001
Trigger	up	-0.001	-0.001
	down	0.001	0.001
PDF	up	0.001	0.001
	down	-0.001	-0.001
W+j kfactor	up	0.009	0.006
	down	-0.009	-0.006
Statistics	up	0.014	0.008
	down	-0.014	-0.008
total	up	0.022	0.026
	down	-0.022	-0.026

[14] arXiv:hep-ph/0603175v2

[15] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," PoS A CAT 040 (2007) [physics/0703039].

[16] Henri Bachacou, A Measurement of the Production Cross Section of tt-bar Pairs Using Secondary Vertex b-tagging, FERMILAB-THESIS-2004-48

Table 2: The full list of systematic uncertainties for the mistag rate scale factor from the $\sigma(t\bar{t})$ method measurement. This uncertainty must be combined with the electron method scale factor uncertainty and the two should be treated as uncorrelated. The uncertainties shown below are absolute shifts, and thus should be applied as $SF \pm shift$, e.g. 1.331 ± 0.094.

mistag SF $\sigma(t\bar{t})$ method		HOBIT Operating Point	
source		Loose	Tight
$\sigma(t\bar{t})$	up	0.007	0.090
	down	-0.007	-0.090
luminosity	up	0.004	0.055
	down	-0.004	-0.055
jet energy scale	up	0.003	0.037
	down	-0.003	-0.037
generator	up	0.003	0.005
	down	-0.003	-0.005
generator	up	0.002	0.023
	down	-0.002	-0.023
ISR/FSR	up	0.000	0.005
	down	-0.000	-0.005
$t \to Wb$ branching ratio	up	0.000	0.005
	down	-0.000	-0.005
Trigger	up	0.000	0.005
	down	-0.000	-0.005
PDF	up	0.000	0.005
	down	-0.000	-0.005
W+j kfactor	up	-0.091	-0.135
	down	0.055	0.081
Statistics	up	0.024	0.125
	down	-0.024	-0.125
total	up	0.094	0.217
	down	-0.060	-0.180

[17] D. Acosta, et al., Measurement of the $t\bar{t}$ Production Cross Section in $p\bar{p}$ Collisions at $\sqrt{=}$1.96 TeV Using Lepton+Jets Events with Semileptonic B Decays to Muons, Phys. Rev. D 71, 052003 (2005)

[18] Nazim Hussain, A simultaneous measurement of the b-tagging efficiency scale factor and the $t\bar{t}$ Production Cross Section at the Collider Detector at Fermilab, FERMILAB-MASTERS-2011-02

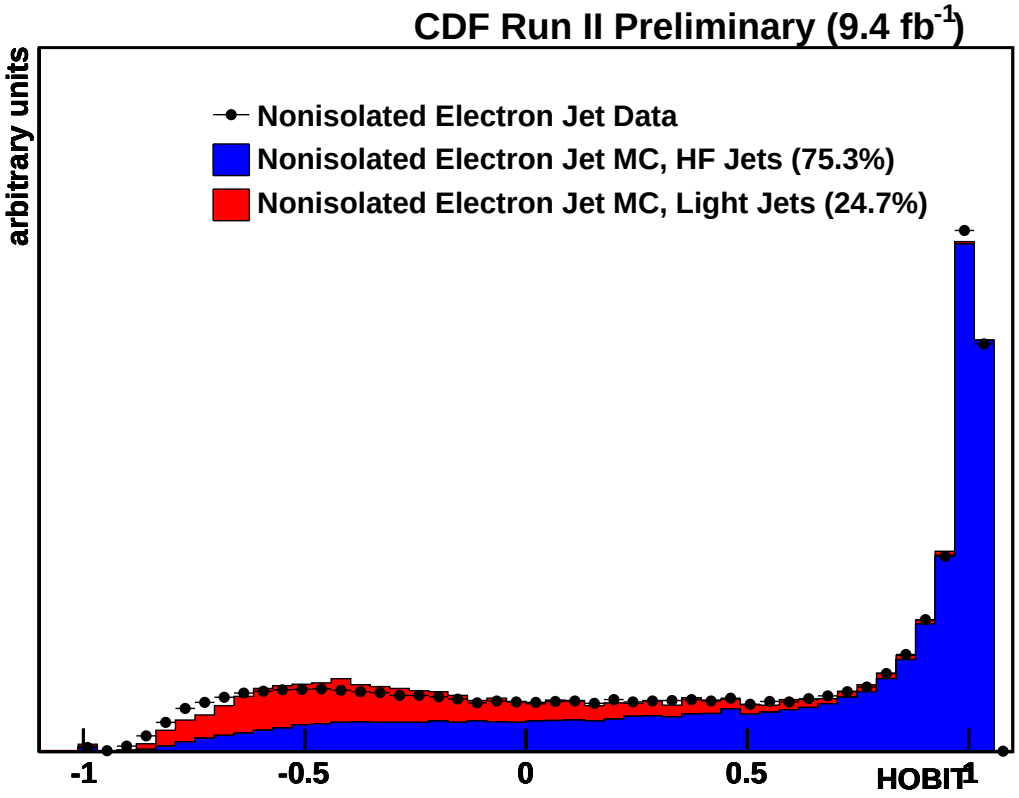

Figure 15: HOBIT output for electron jets, data vs. MC. Relative proportions of HF to light jets are determined via a fit of the two MC templates to the data.

[19] U. Langenfeld, S. Moch and P. Uwer, Measuring the running top-quark mass, Phys. Rev. D 80, 054009 (2009)

[20] T. Aaltonen *et al.* [CDF Collaboration], First Measurement of the Ratio $\sigma_(t\bar{t})/\sigma_(Z/\gamma** \to \ell\ell)$ and Precise Extraction of the $t\bar{t}$ Cross Section, Phys. Rev. Lett. **105**, 012001 (2010) [arXiv:1004.3224 [hep-ex]].

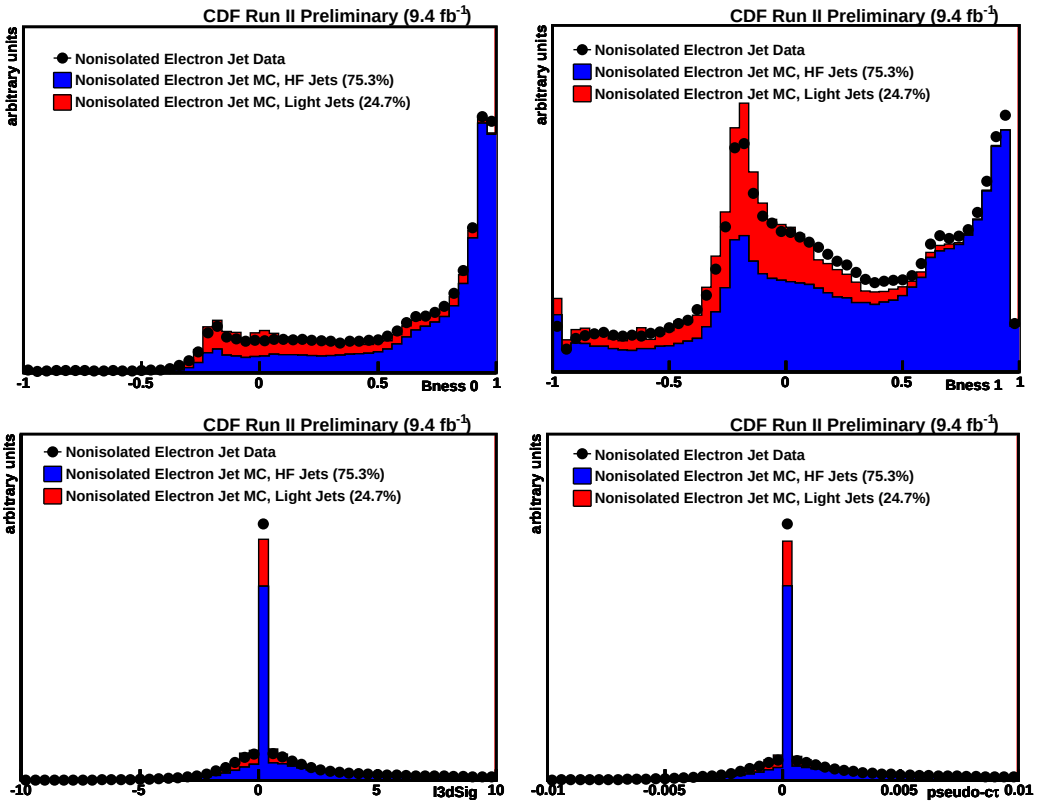

Figure 16: Comparison of select HOBIT inputs for electron jets, data vs. MC.

Table 3: The full list of systematic uncertainties for the b-jet tagging efficiency scale factor from the electron method measurement. This uncertainty must be combined with the $\sigma(t\bar{t})$ method scale factor uncertainty and the two should be treated as uncorrelated. The uncertainties shown below are absolute shifts, and thus should be applied as $SF \pm shift$, e.g. 0.993 ± 0.023.

b-eff SF electron method		HOBIT Operating Point	
source		Loose	Tight
over eff.	up	0.009	0.014
	down	-0.009	-0.014
prescale coor.	up	0.001	0.011
	down	-0.001	-0.011
Et depend.	up	0.010	0.003
	down	-0.010	-0.003
semi-lep bias	up	0.010	0.006
	down	-0.010	-0.006
charm model	up	0.001	0.002
	down	-0.001	-0.002
Stats	up	0.016	0.018
	down	-0.016	-0.018
total	up	0.023	0.026
	down	-0.023	-0.026

Table 4: The full list of systematic uncertainties for the mistag rate scale factor from the electron method measurement. This uncertainty must be combined with the $\sigma(t\bar{t})$ method scale factor uncertainty and the two should be treated as uncorrelated. The uncertainties shown below are absolute shifts, and thus should be applied as $SF \pm shift$, e.g. 1.331 ± 0.092.

b-eff SF electron method		HOBIT Operating Point	
source		Loose	Tight
over eff.	up	0.024	0.092
	down	-0.024	-0.092
prescale coor.	up	0.010	0.003
	down	-0.010	-0.003
Et depend.	up	0.014	0.018
	down	-0.014	-0.018
semi-lep bias	up	0.040	0.055
	down	-0.040	-0.055
charm model	up	0.001	0.004
	down	-0.001	-0.004
Stats	up	0.078	0.163
	down	-0.078	-0.163
total	up	0.092	0.196
	down	-0.092	-0.196